Knowledge-based biomedical word sense disambiguation: comparison of approaches

Antonio Jimeno-Yepes, Alan R. Aronson

Abstract

Word sense disambiguation (WSD) algorithms select the proper sense of ambiguous terms. Resources like the UMLS provide a reference thesaurus to be used to annotate the biomedical literature. Statistical learning approaches have produced good results but the size of the UMLS makes the production of training data infeasible to cover all the domain. We present research on existing WSD approaches based on knowledge-bases, which complement the studies performed on statistical learning.

We compare three approaches. The first approach builds a vector from the definitions, synonyms and related terms for each sense. This vector is compared to the context vector of occurrences of the ambiguous term in text. The second approach builds a query with monosemous synonyms and related terms which is used to retrieve MEDLINE documents, a machine learning approach is trained on this corpus. The last approach exploits the structure of the UMLS Metathesaurus network of relations as a feature to bias the selection of concepts; similar to the Page Rank algorithm.

We have found that building a corpus automatically from MEDLINE using the UMLS Metathesaurus provides better results compared to the other methods. The structure of the UMLS network used to estimate the relevance of the UMLS concepts does not provide a good performance. In addition, the combination of methods improves the performance of individual approaches. On the other hand, the performance is still bellow statistical learning trained on manually produced data and below the maximum frequency sense baseline.

Finally, we propose several directions to improve the existing methods and to improve the UMLS Metathesaurus to be more effective in WSD.

1 Introduction

Word sense disambiguation (WSD) algorithms select the proper sense of ambiguous terms. WSD is an intermediary step within information retrieval and information extraction. In the biomedical domain, interest has been focused mainly on specific entity types (e.g. genes and diseases).

The UMLS Metathesaurus® is the largest biomedical thesaurus available of medical terms collected from more than 100 resources. Several efforts exist to

map the UMLS® to text, (e.g. MetaMap[4] and Whatizit[16]). UMLS 2009AB has at least 24k ambiguous terms, i.e. where a given term is part of more than one concept unique identifier (CUI) in the UMLS Metathesaurus. These ambiguous cases increase if we consider term variability introduced by matching algorithms, where simple morpho-syntactic transformations increase the number of ambiguity cases (e.g. repairing, repaired, repair). All UMLS concepts are assigned one or more broad categories called semantic types. Table 1 shows the distribution of types with the most cases of ambiguity. We can see that proteins, genes and clinical terms are the most ambiguous cases.

Frequency	Type	Description
8,688	T028	Gene or Genome
4,089	T116	Amino Acid, Peptide, or Protein
3,534	T201	Clinical Attribute
2,189	T200	Clinical Drug
1,969	T047	Disease or Syndrome
1,691	T123	Biologically Active Substance
1,408	T170	Intellectual Product
1,278	T121	Pharmacologic Substance
1,252	T126	Enzyme
1,218	T129	Immunologic Factor

Table 1: Top 10 most ambiguous semantic types

Table 2 presents the top terms in MEDLINE® from the ambiguous UMLS terms. Compared to the semantic type statistics, it seems that other semantic types than proteins or genes seem to cover the most interesting cases. We observe that terms like *study* can already be mapped to 6 concepts, showing the complexity of the UMLS content.

In addition, special care is required preprocessing the UMLS since some terms provided by the constituent vocabularies provide some terms like general English terms and numbers which might be difficult to deal with and which might not be of interest in the biomedical context (all, other, had, can, ...). Several procedures [15, 5, 6] have already been studied to perform a cleanup of these cases.

2 Related work

We are interested in performing WSD and to cover as much of the UMLS concepts as possible to improve the MetaMap annotation. Usually, techniques developed using statistical learning have a better performance compared to techniques based on knowledge-based approaches[17]. On the other hand, building a manually annotated corpus, as required by statistical learning, to cover all the concepts in the UMLS Metathesaurus is expensive and infeasible. State of the art knowledge-based approaches rely on graph theory[2] which seems to have an

Frequency	Term	Amb. level
3,215,158	study	6
2,122,371	treatment	4
2,064,598	all	6
1,955,592	2	5
1,945,251	1	5
1,872,536	other	44
1,795,137	had	2
1,762,387	effect	2
1,757,672	can	11
1,755,725	cell	4

Table 2: Top 10 most ambiguous terms in MEDLINE

interesting performance but still far from the maximum frequency sense baseline or the statistical learning approaches.

Previous work in WSD for the UMLS includes knowledge-based and supervised methods. Among the knowledge-based methods we find the Journal Descriptor Indexing method[8] and several based on graph methods [3]. Machine learning algorithms have been explored in several studies where alternative combinations of features are compared[11, 19, 3, 13], these studies obtain a performance over 0.86 in terms of accuracy.

Related work in the biomedical domain shows that statistical learning performs better than unsupervised or knowledge based ones. Existing corpora in the biomedical domain[20, 3] cover just a small number of terms and senses compared to the content of the UMLS Metathesaurus. Extending manually existing corpora to cover the UMLS Metathesaurus does not seem to be feasible.

The idea of generating corpora automatically to perform WSD has already been presented in the WSD literature. Leacock et al.[12] used monosemous relatives and co-occurrences to retrieve training data. Their automatically generated dataset showed promising results but not as good as training with manually generated data. Agirre and Martinez[1] built corpora for WSD based on the Web. In their work, evaluated on Senseval, they show the feasibility of building such a corpus and better results are obtained on corpus biased following the sense distribution.

The automatic acquisition of corpora to perform WSD has already been successfully used in the biomedical domain to disambiguate acronyms[7]. In this case, the occurrences of long forms and acronyms are located using pattern matching. The examples are collected and processed to perform learning based on the SVM learning algorithm.

Machine learning approaches seem to obtain better results than unsupervised or knowledge based methods. In this paper, we compare several knowledge-based methods not based training data on the UMLS Metathesaurus concepts. In the following section, we present the automatic preparation of training data from MEDLINE based on the UMLS Metathesaurus

3 Methods

We compare three knowledge-based methods which have different assumptions on the context of the terms and how the terms should be disambiguated. The first method compares a profile vector of the UMLS Metathesaurus concept which is compared to the context vector of the ambiguous term. The second method complements the terms in the context with statistics from the UMLS Metathesaurus network of relations, which are used to influence the decision of the disambiguation algorithm. The third method looks for training data collected automatically using PUBMED® queries built out of the monosemous synonyms and related terms of the senses of the ambiguous term. The retrieved documents are used to train a Naïve Bayes classifier.

3.1 Machine Readable Dictionary context

In the first approach, the context of the words surrounding the ambiguous word is compared to a profile built from a UMLS concept which includes the definition, synonyms and related terms. This algorithm can be seen as a relaxation of Lesk's algorithm [14], which is very expensive since the sense combination might be exponentially large even for a single sentence. The literature has shown that similar or even better performance might be obtained disambiguating each ambiguous word separately.

A concept profile vector has as dimensions the tokens in the definition, synonyms and related terms. Stop words are discarded and the Porter stemming is used to normalize the tokens. In addition, the token frequency is normalized based on the inverted *concept* frequency. This means that terms which are repeated many times within the UMLS will have less relevance.

A context vector for an ambiguous term includes the term frequency, stop words are removed and porter stemmer is applied. The word order is lost in the conversion.

In the machine readable approach (MRD), vectors of concept profiles c linked to an ambiguous word w in set C_w and word contexts cx are compared using the cosine similarity as shown in equation 1; the concept with the highest cosine similarity is selected.

$$MRD(c) = \underset{c \in C_w}{\operatorname{argmax}} \frac{c \cdot cx}{|c||cx|} \tag{1}$$

3.2 Page Rank WSD implementation

This second approach combines the context of the word with the chances of selecting the concept based on the topology of the network of the resource used for disambiguation. The algorithm was developed by Agirre and Sorao[2]. It is inspired by the Google Page Rank algorithm, which is used to encode word sense dependencies using random walks on graphs.

In this approach the knowledge resource is represented as follows. Let G be a graph with N vertices $v_1, ..., v_N$, d_i be the outdegree of node i; let matrix

M be an $N \times N$ transition probability matrix. $M_{ij} = \frac{1}{d_i}$. To estimate the PageRank vector Pr over G, requires solving equation 2, where v is an $N \times 1$ vector of elements $\frac{1}{N}$ and c is a *damping factor*.

$$Pr = cMPr + (1 - c)v \tag{2}$$

The UMLS Metathesaurus has been processed as follows to prepare it for the PageRank algorithm. The terms in the UMLS Metathesaurus are normalized using the Porter stemmer. Spaces are replaced by underscore characters. A dictionary file (containing terms and pointers to concepts) and a relation file are produced according to Agirre and Sorau's implementation.

The context of an ambiguous word defined by words within a specified window is tokenized, stopwords are removed and Porter stemming is applied. Tokens in the context and in the dictionary file might not match in some cases since UMLS Metathesaurus terms might contain multiple words. Tokens from the context are finer grained than the UMLS terms.

In competitions like Senseval[1], where this technique has been previously evaluated, term boundaries are properly specified. Manual processing of the corpus is not possible in our case due to its large size (e.g. MEDLINE). Automatic processing of text might help to get the text ready. Named entity recognition might be used but there is no training data to define term boundaries. Processing with NLP tools could harm the results due to term misalignments.

This means that Agirre and Sorau's tool might lack recall. The limitation of using this tool in this environment has to be properly understood. The tool will be mainly used to study the influence of the UMLS network of relations on deciding the best sense; it complements the other two approaches.

3.3 Automatic corpus extraction from MEDLINE

In this third approach, corpora to train for statistical learning algorithms for ambiguous terms are prepared retrieving documents from a corpus.

The UMLS Metathesaurus is used to obtain information related to the candidate concepts linked to an ambiguous term. We use MEDLINE[2] as our corpus, which is the largest bibliographic database in the biomedical domain with citations from around 5,000 journals.

Queries are generated using English monosemous relatives of the candidate concepts which, potentially, have an unambiguous use in MEDLINE. We have performed experiments using synonyms which were not ambiguous in the UMLS. In addition, we have used the ambiguous term combined with unambiguous related terms, which might occur in MEDLINE, assuming one sense per document. Long terms are discarded since they might not appear in MEDLINE and very short terms and numbers are discarded to avoid ambiguous terms. A standard stop word list is used to remove uninformative English terms.

[1] http://www.senseval.org

[2] http://www.nlm.nih.gov/databases/databases_medline.html

CUI: C0374711

```
"Surgical repair"[tiab]
OR ("repair"[tiab] AND
                       ("Corneal Transplantation"[tiab]
                     OR "Corneal Transplantations"[tiab]
                     OR "Corneal Graftings"[tiab]
                     OR "Corneal Grafting"[tiab]
                     OR "Cornea Transplantations"[tiab]
...
                     OR "Repair of the Middle Ear"[tiab])
   )
```

CUI:C0043240

```
"Wound Healings"[tiab] OR "Wound Repair"[tiab]
OR ("repair"[tiab] AND
                       ("Granulation Tissues"[tiab]
                     OR "Natural regeneration"[tiab]
                     OR "Blood Clottings"[tiab]
                     OR "BLOOD COAG"[tiab]
                     OR "COAG BLOOD"[tiab]
...
                     OR "Integrin alphaIIbbeta3"[tiab])
   )
```

Figure 1: Query example for term *repair* using synonyms and related concepts

We have used EUtils[3] from PUBMED[4] as the search engine to retrieve documents from MEDLINE. In order to retrieve documents where the text (title or abstract of the citation) contains the query terms, the "[tiab]" search field is used. Quotes are used to find exact mentions of the terms and increase precision. Examples of queries for the ambiguous term *repair*, with concept identifiers *C0374711* and *C0043240*, using monosemous relatives are found in figure 1.

Documents retrieved using PUBMED are assigned to the concept which was used to generate the query. If no documents are returned, the quotes are replaced by parentheses to allow finding the terms in any position in the title or abstract text. We have evaluated several limits on the number of retrieved documents. Since there is not a significant difference, 100 documents are collected from MEDLINE for each concept identifier.

The automatically generated corpus is used with Naïve Bayes (NB). The trained model is then evaluated against already annotated sets from where precision and recall values are recorded as shown in the results section.

[3] http://eutils.ncbi.nlm.nih.gov/

[4] http://www.ncbi.nlm.nih.gov/pubmed

4 Results

The NLM WSD data set[20] has been used to conduct the experiments. This set contains 50 ambiguous terms and annotations of UMLS semantic types. In addition, there is a mapping to the UMLS unique concept identifiers (CUI) for the 1999 version. If there is no UMLS concept identified in the text, *None of the above* has been assigned in the NLM WSD.

We have considered the same setup as Humphrey et al.[8] and discarded the *None of the above* category. As the ambiguous term *association* has been assigned entirely to *None of the above*, it has been discarded. This means that we will present results for 49 out of the 50 ambiguous terms.

Results are presented in table 3 in terms of weighted average precision, recall and F-measure. Table 3 shows the results comparing the use of monosemous synonyms, related terms with machine learning (MS-RT-NB), machine readable dictionary (MRD), Page Rank (PPR) and several baselines. We show, as well, a variant of the maximum frequency sense where the frequencies are obtained from the queries using monosemous synonyms and related terms (MFS_Medline).

Words occurring in the abstract, where the ambiguous terms appear, are used as the context of the ambiguous word. All three algorithms have used this context to perform disambiguation.

We have used several baselines which allow comparing different assumptions. One baseline is maximum frequency sense (MFS), which is standard in WSD evaluation. The other baseline is statistical learning based on Naïve Bayes and the NLM WSD set; 10-fold cross-validation sampling is used.

1999	Precision	Recall	F-Measure
MRD	0.8532	0.6350	0.7281
PPR	0.6700	0.5727	0.6175
MS + RT + NB	**0.8673**	**0.6836**	**0.7646**
Comb. linear	0.8675	0.6923	0.7701
Combine vote	0.8581	0.7045	0.7738
MFS	0.7460	0.8637	0.8005
NB	**0.8714**	**0.8863**	**0.8788**

Table 3: NLM WSD results: method comparison

Machine learning on the NLM WSD set has the best performance in terms of precision and recall, showing better performance than the MFS baseline as already shown in the literature. MFS indicates that usually one sense of the term is highly represented compared to the rest of the senses. These two baselines require special consideration since we cannot know which is the sense with the highest frequency or have training data to train a classifier to perform disambiguation.

The use of monosemous synonyms and related terms performs better than the other knowledge-based methods. On the other hand, the results are below the MFS baseline in terms of recall and F-measure but higher in terms of pre-

cision. The machine readable dictionary approach produces results not as good compared to automatically produced training corpus. The page rank approach presents the lowest performance compared to any of the approaches presented in this study.

The combination of the approaches has been done either by maximum vote (Combine vote) or by linear combination of the prediction probability or score assigned to the senses (Comb. linear). All the knowledge-based WSD approaches provide a numerical value between 0 and 1. Table 3 indicates that there is an improvement of the performance; but it is not large compared to the best performing method.

Considering time performance, the MRD approach just takes a couple of minutes for all the cases available for the 49 ambiguous terms. Retrieving, building the classifier and classifying the test cases is a bit more expensive than the MRD approach. But if we just use the trained classifier, the speed is faster than the MRD approach. Page Rank approach is by far the most expensive of the methods and it took several hours to disambiguate all the cases. Agirre and Soroa already stated this in their experiments using WordNet.

5 Discussion

The automatically extracted corpus seems to produce better performance than the MRD and PPR approach. There are several possible explanations. The MRD relies on the terms presented in the dictionary, in this case the UMLS dictionary. We identify related terms but in some cases these terms are not representative of the context for a given sense. On the other hand, the automatic extracted corpus seems to rely on the UMLS content to collect documents from MEDLINE which might expand the context terms and, in addition, rely on statistical learning approaches which might produce a better partition of the feature space.

Among the best performing terms with the automatically extracted corpus we find: *transport*, *support*, *resistance*, *depression* and *strains*. These cases are homonyms (not polysemous); so their senses are easy to identify. On the other hand, the terms with the lowest performance are: *growth*,*determination*, *surgery*, *nutrition* and *blood pressure*. In these cases, the differences are blurred and seem to be closer in meaning. An exception is *growth*, the terms in the UMLS 1999 for the M2 (*Functional Concept*) sense are contained in the M1 (*Organism Function*) sense. The term *blood pressure* could indicate the measurement the blood pressure or the blood pressure of a patient. These senses are difficult to distinguish and the UMLS did not provide enough evidence to provide a retrieval query specific enough make the distinction.

The term *cold* has a low performance as well. This is strange since the annotators found that the senses for *cold* were clearly distinct in the documents. Looking at the confusion matrix of the Naïve Bayes classifier, we have found that a large proportion of instances belonging to the sense M1 (*Cold Temperature*) have been classified as M5 (*Cold Sensation*). Looking at the retrieved

documents we find as well that one of the assumptions by Yarowsky (one sense per document) does not hold in the *cold* case since multiple meanings of the term cold happen in the documents retrieved for the sense M5. Further refinement of the terms in the UMLS Metathesaurus might retrieve better documents since terms for *cold temperature* did not retrieve some of the documents. In addition, disambiguation approaches looking closer at the word context instead of the abstract level might improve the results.

The automatically generated queries are not specific enough in some cases, so they retrieve false positives for a given sense. The results are in tune with general English results, where the performance is lower than using manually generated training data. We identify similar cases where senses are not so clearly distinct. On the other hand, these cases are more difficult to spot from text compared to similar tasks in the biomedical domain where acronyms are the ambiguous terms to disambiguate[7] and the long form is used to identify the correct term. Specific needs for WSD could be studied with these techniques. Identifying further heuristics for a more general disambiguation approach is welcome.

Among the terms for which the MRD approach has the best performance we find *depression*, *determination*, *transport*, *strains* and *transient*. Some of these terms match the ones from the automatically extracted corpus. In the case of *transport*, in the biological sense terms like *process* or *metabolism* are within the most relevant terms and in the patient sense we find *patient* and *delivery*. The context defined by the concept vectors allows properly differentiating the sense in text.

On the other hand there are some cases in which the MRD approach cannot properly disambiguate properly. Among these cases we find: *single*, *scale*, *nutrition*, *adjustment* and *man*. Despite the fact that some of the terms might be confusing in context (e.g. *man*), in these cases, the concept profiles might not be representative of the ambiguous term senses. So, the terms with higher $tf \times idf$ are not representative of the context of the ambiguous words.

Considering the term *scale*, the sense M2 (*Intellectual Scale*) is the preferred one, in most of the cases the M1 (*Integumentary Scale*) is selected. Looking at the concept profiles in table 4, we find that the terms in M2 do not really seem to contain terms which could co-occur with scale in the M2 context. In addition, the vector for M1 is very short, containing two dimensions (*integumentary*, *scale*), so matching the term *scale* biases the sense prediction to M1.

In the case of *nutrition*, which also has low performance in the noisy-corpus approach, we find that the vectors have similar terms with high $tf \times idf$ (c.f. table 5). In the WSD results, we find that the correct senses are split among M1 (*Organism Attribute*) and M3 (*Feeding and dietary regimes*) and no ambiguous term is assigned to M2 (*Science of Nutrition*). The MRD approach classifies M3 cases as M2 or M1 and some M1 cases are assigned to M2. No annotation is done to M3.

The cosine method defines a point in the space for the concept from which the distance is estimated. This means that the feature space would look like a sphere, being the center the vector of the concept profile and the maximum radius will match the space of the following concept. This space might be

M1		M2		M3	
Term	$tf \times idf$	Term	$tf \times idf$	Term	$tf \times idf$
scale	19.06	scale	68.75	scale	55.30
integumentari	8.39	interv	25.17	weight	46.06
		seri	24.74	measur	41.91
		loinc	22.52	compon	33.80
		sequenc	21.38	devic	31.98

Table 4: Scale top $tf \times idf$ terms for the senses M1, M2, M3

M1		M2		M3	
Term	$tf \times idf$	Term	$tf \times idf$	Term	$tf \times idf$
nutrit	1519.81	nutrit	1318.84	nutrit	158.13
physiolog	548.57	scienc	453.13	scienc	81.95
avail	205.97	health	433.43	statu	35.07
statu	182.38	physiolog	351.14	regim	13.48
phenomena	131.35	food	311.01	outcom	10.17

Table 5: Nutrition top $tf \times idf$ terms for the senses M1, M2, M3

restrictive for some senses and might explain some of the better performance of the automatically extracted corpus approach. Statistical learning approaches might be capable of defining a more detailed feature space. Similar conclusions have been drawn for similar methods in text categorization[18].

The page rank approach (PPR) has inferior performance compared to the other methods. One possible reason is the assumption that concepts with larger number of related concepts might be more relevant does not hold for the UMLS. The terms with the best performance have a large number of relations linked to the right sense, this means that there is a large number of sense linked to that sense. These terms are: *transient*, *scale*, *reduction*, *frequency* and *fit*. The same applies for the terms with the worst performance: *determination resistance inhibition transport* and *sensitivity*. This implies that the number of relations in the UMLS does not directly imply relevance of the sense.

6 Conclusions and Future Work

We have compared several methods for knowledge-based sense disambiguation in the biomedical domain. We find that an automatic extracted corpus used to train statistical learning approaches has the best performance. On the other hand, these methods do not achieve as good performance as the maximum frequent sense baseline or statistical learning approaches trained on manually generated training data.

Machine readable approaches seem to have a lower performance than the automatically extracted corpus. This seems to be due to the inadequacy of the UMLS for the task in some cases; it is not the purpose of the UMLS to perform WSD and we can foresee some research to produce a UMLS version

tuned for WSD. In addition, the context based distance approaches might require some corpus statistics as the ones obtained with the automatic extracted corpus to complement them; it might be difficult to identify a proper distance measurement which is appropriate for all the cases.

Semantic network metrics have shown to be perform less well than other methods. This means that a higher chance of selecting a sense from the dictionary does not necessarily imply relevance.

The combination of the predictions of the methods used here performs better than any individual method. Even though the increase is not large, it shows that each method has a complementary view on the data.

Automatic extraction of a corpus from MEDLINE seems to provide good results but still has some drawbacks. Filtering of documents to improve the quality of the automatically extracted corpus could improve the performance of the statistical learning algorithms on the automatically extracted corpus.

Polysemous terms have been difficult to cope with. The feature set (tokens) might not provide enough detail. Additional set of features might be useful in providing further input to the different algorithms. UMLS concepts are assigned semantic types, it might be possible to automatically obtain a higher level categorization which might be linked to the UMLS categorization. In this scenario, we could explore disambiguation scenarios where a narrower purpose is defined, e.g. gene normalization, where specific heuristics can be applied and complement.

Corpus statistics might help to complement the UMLS and improve WSD methods or related text mining tasks. For example, corpus statistics might help to optimize the UMLS Metathesaurus to improve the document retrieval from MEDLINE. Several ideas have already been proposed to clean up an existing thesaurus[10, 5, 6] and to add further relevant content[9].

Knowledge-based approaches have good performance, even though below standard WSD baselines. We have presented several approaches and analyzed their performance and drawbacks. Finally, we have proposed several directions for further research which might improve their performance, and some of them could be used to improve the UMLS for WSD.

References

[1] E. Agirre and D. Martinez. Unsupervised WSD based on automatically retrieved examples: The importance of bias. In *Proceedings of EMNLP*, pages 25–32, 2004.

[2] E. Agirre and A. Soroa. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41. Association for Computational Linguistics, 2009.

[3] D. Alexopoulou, B. Andreopoulos, H. Dietze, A. Doms, F. Gandon, J. Hakenberg, K. Khelif, M. Schroeder, and T. Wächter. Biomedical word sense

disambiguation with ontologies and metadata: automation meets accuracy. *BMC bioinformatics*, 10(1):28, 2009.

[4] A. Aronson and F. Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229, 2010.

[5] A. Aronson, J. Mork, A. Névéol, S. Shooshan, and D. Demner-Fushman. Methodology for Creating UMLS Content Views Appropriate for Biomedical Natural Language Processing. In *AMIA Annual Symposium Proceedings*, volume 2008, page 21. American Medical Informatics Association, 2008.

[6] D. Demner-Fushman, J. Mork, S. Shooshan, and A. Aronson. UMLS content views appropriate for NLP processing of the biomedical literature vs. clinical text. *Journal of Biomedical Informatics*, 2010.

[7] S. Gaudan, H. Kirsch, and D. Rebholz-Schuhmann. Resolving abbreviations to their senses in Medline. *Bioinformatics*, 21(18):3658, 2005.

[8] S. Humphrey, W. Rogers, H. Kilicoglu, D. Demner-Fushman, and T. Rindflesch. Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: Preliminary experiment. *Journal of the American Society for Information Science and Technology (Print)*, 57(1):96, 2006.

[9] A. Jimeno-Yepes, R. Berlanga-Llavori, and D. Rebholz-Schuhmann. Ontology refinement for improved information retrieval. *Information Processing & Management*, 2009.

[10] A. Jimeno-Yepes, R. Berlanga-Llavori, and D. Rebholz-Schuhmann. Terminological cleansing for improved information retrieval based on ontological terms. In *Proceedings of the WSDM'09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 6–14. ACM, 2009.

[11] M. Joshi, T. Pedersen, and R. Maclin. A comparative study of support vector machines applied to the supervised word sense disambiguation problem in the medical domain. In *Proceedings of the 2nd Indian International Conference on Artificial Intelligence (IICAI05)*, pages 3449–3468. Citeseer, 2005.

[12] C. Leacock, G. Miller, and M. Chodorow. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–165, 1998.

[13] G. Leroy and T. Rindflesch. Effects of information and machine learning algorithms on word sense disambiguation with small datasets. *International Journal of Medical Informatics*, 74(7-8):573–585, 2005.

[14] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM, 1986.

[15] J. Mork and A. Aronson. Filtering the UMLS Metathesaurus for MetaMap. Technical report, Tech rep, National Library of Medicine, 2009.

[16] D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch, and A. Jimeno. Text processing through Web services: calling Whatizit. *Bioinformatics*, 24(2):296, 2008.

[17] M. Schuemie, J. Kors, and B. Mons. Word sense disambiguation in the biomedical domain: an overview. *Journal of Computational Biology*, 12(5):554–565, 2005.

[18] F. Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[19] M. Stevenson, Y. Guo, and R. Gaizauskas. Acquiring sense tagged examples using relevance feedback. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 809–816. Association for Computational Linguistics, 2008.

[20] M. Weeber, J. Mork, and A. Aronson. Developing a test collection for biomedical word sense disambiguation. In *Proceedings of the AMIA Symposium*, page 746. American Medical Informatics Association, 2001.